LINGUISTIC SEGMENTATION OF SPEECH

RELATED APPLICATIONS

[0001] This application claims priority under 35 U.S.C. § 119 based on U.S. Provisional Application Nos. 60/394,064 and 60/394,082 filed July 3, 2002 and Provisional Application No. 60/419,214 filed October 17, 2002, the disclosures of which are incorporated herein by reference.

GOVERNMENT INTEREST

[0002] The U.S. Government has a paid-up license in this invention as provided by the terms of contract No. N66001-00-C-8008 awarded by the Defense Advanced Research Projects Agency (DARPA).

BACKGROUND OF THE INVENTION

A. Field of the Invention

[0003] The present invention relates generally to speech processing and, more particularly, to linguistic segmentation of transcribed speech.

B. Description of Related Art

[0004] Speech has not traditionally been valued as an archival information source. As effective as the spoken word is for communicating, archiving spoken segments in a useful and easily retrievable manner has long been a difficult proposition. Although the act of recording audio is not difficult, automatically

transcribing and indexing speech in an intelligent and useful manner can be difficult.

[0005] Speech is typically received into a speech recognition system as a continuous stream of words. In order to effectively use the speech in information management systems (e.g., information retrieval, natural language processing, real-time alerting), the speech recognition system initially transcribes the speech to generate a textual document. A simple transcription, however, will generally not contain significant information that was present in the original speech. For example, the transcription may be a mere stream of words that lack many of the linguistic features that a listener of the speech would normally identify.

[0006] Linguistic features that may be lacking in a simple transcription of speech include linguistic features that are visible in text, such as periods, quotation marks, exclamation marks, commas, and direct quotation marks. Additionally, linguistic features may include non-visible information, such as phrasal boundaries.

[0007] There is a need in the art to be able to automatically generate linguistic information, including visible and non-visible linguistic features, for audio input streams.

SUMMARY OF THE INVENTION

[0008] Systems and methods consistent with the principles of this invention provide a linguistic segmentation tool that generates a comprehensive set of linguistic information for a document transcribed based on human speech.

[0009] One aspect of the invention is directed to a linguistic segmentation tool. The linguistic segmentation tool includes a lexical feature extraction component configured to receive text and generate lexical feature vectors relating to the text. The linguistic segmentation tool further includes an acoustic feature extraction component that receives a spoken version of the text and generates acoustic feature vectors relating to the spoken version of the text. Finally, the linguistic segmentation tool includes a statistical framework component configured to generate linguistic features associated with the text based on the acoustic feature vectors and the lexical feature vectors.

[0010] A second aspect of the invention is directed to a method for determining linguistic information for words corresponding to a transcribed version of speech. The method includes generating lexical features for the words, including a syntactic class associated with the words and generating acoustic features for the speech. The acoustic features are based on speaker pauses, speaker rate, speaker energy, and/or speaker pitch. The method further includes generating the linguistic information based on the lexical features and the acoustic features.

[0011] Yet another aspect consistent with the invention is directed to a computing device for determining linguistic information for words corresponding to a transcribed version of speech. The computing device includes a processor and a computer memory coupled to the processor and containing programming instructions. The program instructions, when executed by the processor, cause the processor to generate lexical features for the words, including a syntactic

class associated with at least one of the words, and generate acoustic features for the speech. The acoustic features are based on speaker pauses, speaker rate, speaker energy, and/or speaker pitch. The program instructions further cause the processor generate the linguistic information based on the lexical features and the acoustic features, and output the generated linguistic information as meta-information embedded in the transcribed version of the speech.

[0012] Yet another aspect of the invention is a method for associating meta-information with a document transcribed from speech. The method includes building a language model based on lexical feature vectors extracted from the document, where the lexical feature vectors include a word and a syntactic classification of the word. The method further includes building an acoustic model based on acoustic feature vectors extracted from the speech and combining outputs of the language model and the acoustic model in a statistical framework that estimates a probability for associating the meta-information with the document.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate the invention and, together with the description, explain the invention. In the drawings,

[0014] Fig. 1 is a diagram illustrating an exemplary system in which concepts consistent with the invention may be implemented;

[0015] Fig. 2 is a block diagram conceptually illustrating linguistic segmentation;

[0016] Fig. 3 is a block diagram of a linguistic segmentation tool consistent with the present invention;

[0017] Fig. 4 is a diagram illustrating a series of words;

[0018] Fig. 5 is a flow chart illustrating methods for assigning a syntactic class to a word; and

[0019] Fig. 6 is a flow chart illustrating methods for estimating the probability of the occurrence of a linguistic feature.

DETAILED DESCRIPTION

[0020] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers may be used in different drawings to identify the same or similar elements. Also, the following detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents of the claim limitations.

[0021] Linguistic segmentation of spoken audio is performed by a linguistic segmentation tool based on a transcribed version of the speech and the original speech. The linguistic segmentation tool analyzes both lexical and acoustical features of the speech in generating the linguistic segments. The lexical features include syntactic classifications of the words in the transcribed text. The acoustical features include measured pauses, speaking rate, speaker energy, and speaker pitch. Speech models based on the acoustic and lexical features

are combined to achieve a final probability of a particular linguistic feature occurring.

SYSTEM OVERVIEW

[0022] Linguistic segmentation, as described herein, may be performed on one or more processing devices or networks of processing devices. Fig. 1 is a diagram illustrating an exemplary system 100 in which concepts consistent with the invention may be implemented. System 100 includes a computing device 101 that has a computer-readable medium 109, such as random access memory, coupled to a processor 108. Computing device 101 may also include a number of additional external or internal devices, such as, without limitation, a mouse, a CD-ROM, a keyboard, and a display.

[0023] In general, computing device 101 may be any type of computing platform, and may be connected to a network 102. Computing device 101 is exemplary only. Concepts consistent with the present invention can be implemented on any computing device, whether or not connected to a network.

[0024] Processor 108 can be any of a number of well-known computer processors, such as processors from Intel Corporation, of Santa Clara, California. Processor 108 executes program instructions stored in memory 109.

[0025] Memory 109 contains an application program 115. In particular, application program 115 may implement the linguistic segmentation tool described below. The linguistic segmentation tool 115 may receive input data, such as linguistically segmented text, from other application programs executing

in computing device 101 or in other computing devices, such as those connected to computing device 101 through network 102. Linguistic segmentation tool 115 processes the input data to generate indications of linguistic features.

LINGUISTIC SEGMENTATION TOOL

[0026] Fig. 2 is a block diagram conceptually illustrating linguistic segmentation as performed by linguistic segmentation tool 115. An audio input stream having speech is initially processed as a lexical model 201 and an acoustic model 202. Lexical model 201 operates on a number of lexical feature vectors that describe lexical features of the input speech. Acoustic model 202 operates on prosodic features, such as, for example, pauses, speaker rate, energy, and pitch.

[0027] The prosodic features and the lexical features are combined by statistical framework 203 to generate the linguistic features. The linguistic features may be generally referred to as meta-information that further defines meaning beyond the plain words in the input speech. The meta-information may include visible meta-information that is normally associated with written documents, such as period marks, quotation mark, exclamation marks, commas, and direct quotation marks. Additionally, meta-information that is invisible to the traditional written document, such as phrasal boundaries and structured speech locations, may also be generated by statistical framework 203. The complete output document, including the plain words of the document and the linguistic meta-information, is output by statistical framework 203 and may be used by

other information management systems (e.g., information retrieval and natural language processing systems) that add value to the archived speech.

[0028] Fig. 3 is a block diagram illustrating elements of linguistic segmentation tool 115 in additional detail. Segmentation tool 115 includes a speech recognition system 301, an acoustic feature extraction component 302, and a statistical framework component 303.

[0029] Speech recognition system 301 includes transcription component 310 and lexical feature extraction component 311.

[0030] Transcription component 310 receives the input audio stream and generates a textual transcription of the audio stream. Transcription component 310 may be an automated transcription tool or its output text may be generated through a manual transcription process. The output of transcription component 310 is received by lexical feature extraction component 311. Lexical feature extraction component 311 generates the lexical feature vectors that describe the lexical features of the input speech (described in more detail below).

[0031] Acoustic feature extraction component 302 generates acoustic feature vectors based on the input audio information (described in more detail below). The acoustic feature vectors describe prosodic information from the input audio.

[0032] Statistical framework component 303 receives the lexical and acoustic vectors from lexical feature extraction component 311 and acoustic feature extraction component 302, respectively, and based on these vectors, generates a language model (LM) 315 and an acoustic model (AM) 316 for the speech. Statistical framework component 303 combines the outputs of these models to

generate the final lexical features. Statistical framework component 303 may output a linguistically segmented document, which includes the originally transcribed text with meta-information describing the linguistic features.

ACOUSTIC FEATURE EXTRACTION

[0033] Acoustic feature extraction component 302 extracts acoustic feature vectors that correspond to boundaries between words. Fig. 4 is a diagram illustrating a series of words (labeled as words w_1, w_2, w_3, w_4). In one implementation, an acoustic feature vector is generated at each word boundary, labeled as boundaries 401.

[0034] Each acoustic feature can be thought of as a function based on the acoustic information to the left of a particular boundary 401 ($Info_L$) and the acoustic information to the right of a particular boundary 401 ($Info_R$). In other words, an acoustic feature for a particular boundary is defined as

$$f(Info_L, Info_R),$$

where *f* indicates the function. In one implementation, this function may be implemented as a difference operation. In an alternate implementation, this function may be implemented as $\log(Info_L/Info_R)$.

[0035] The information assigned to $Info_L$ and $Info_R$ may be based on four basic prosodic features: (1) speaker pauses (e.g., pause duration), (2) speaker rate (e.g., duration of vowels; either the absolute value of vowel durations or differences in vowel durations), (3) speaker energy (signal energy), and (4) pitch (e.g., absolute pitch values or changes in pitch). In one implementation, function

f is applied 26 times to 26 different combinations of $Info_L$ and $Info_R$ that are selected from prosodic features (1)-(4). In this manner, for each boundary 401 acoustic feature extraction component 302 generates an acoustic feature vector containing 26 acoustic features.

[0036] One of ordinary skill in the art will recognize that in alternate implementations, more or less than 26 acoustic features can be used in a single acoustic vector.

LEXICAL FEATURE EXTRACTION

[0037] Lexical feature extraction component 311 generates lexical feature vectors for the series of words it receives from transcription component 310. A lexical feature vector includes an indication of a word and a syntactic class of the word (described below). Other features may be included in the lexical feature vector, such as the structured speech member of the word (e.g., whether the word is a proper name, a number, an email address, or a URL).

[0038] The syntactic class of a word is an indication of the role of the word relative to its surrounding words. For example, possible syntactic classes may indicate whether a word tends to start a sentence, connect phrases together, or end phrases. In one implementation, potential syntactic classes are automatically generated by lexical feature extraction component 311.

[0039] Fig. 5 is a flow chart illustrating methods for assigning a syntactic class to a word.

[0040] A first set of syntactic classes is based on word affixes. The suffix or prefix of a word often implies the role of the word. Lexical feature component 311 stores a list of word suffixes/prefixes that are known to have a strong probability of implying the role of the word. The list of suffixes/prefixes may be manually generated by a human expert or the list may be automatically learned by feature extraction component 311 from training document(s). In one implementation, for the English language, approximately 30-40 suffixes/prefixes are used, corresponding to 30-40 suffix/prefix classes.

[0041] For each word, lexical feature component 311 begins by determining if the word has a suffix or prefix that matches the predefined list of suffixes/prefixes (act 501). If so, the word is assigned a syntactic class corresponding to its suffix/prefix (act 502).

[0042] In addition to assigning syntactic classes based on suffixes/prefixes, lexical feature component 311 assigns classes based on a predefined set of "function words." The list of function words is based on word frequency. For example, in one implementation, the approximately 2000 most frequently occurring words in a language may be considered function words. If the word being examined by lexical feature component 311 is one of these function words, (act 503), the word is assigned a syntactic class corresponding to the function word (act 504).

[0043] Lexical feature component 311 assigns words that do not have matching suffixes/prefixes and are not function words to an undefined "catch-all" syntactic class (act 505).

[0044] In the manner described above, words are assigned to one of approximately 2030 classes (30 suffix/prefix classes, 2000 function word classes, one catch-all class) by lexical feature component 311. The class assignment, along with the word itself, and possibly along with other lexical features, such as the structured speech member of the word, defines the word's lexical feature vector.

GENERATION OF LINGUISTIC SEGMENTATION INFORMATION

[0045] Statistical framework component 303 receives the acoustic feature vectors from acoustic feature extraction component 302 and the lexical feature vectors from lexical feature component 311. More specifically, statistical framework component 303 may construct a language model 315 (LM) based on the lexical feature vectors and an acoustic model 316 (AM) based on the acoustic feature vectors. The language model and the acoustic model are combined using maximum likelihood estimation techniques to obtain a final probability that a particular one of the linguistic features (e.g., period, exclamation, phrasal boundary, etc.) is present at the location corresponding to the acoustic and lexical feature vector.

[0046] In one implementation, language model 315 is a tri-gram model that estimates the probability of a particular language vector corresponding to a word boundary based on the present language vector and the two previous language vectors. Stated more formally, the language model 315 may be defined as:

$$\text{LM:} \quad P(LV_i|LV_{i-1}, LV_{i-2}),$$

where *P* is the probability of a boundary at the i^{th} language vector (LV) given the previous two language vectors (LV_{i-1} and LV_{i-2}).

[0047] Acoustic model 316 may include an acoustic model that estimates the probability of occurrence of each of the potential linguistic features. In one implementation, acoustic model 316 may include a neural network, such as a three layer neural network having 26 input nodes (one for each acoustic feature in the acoustic vector), 75 hidden layer nodes, and a single output node. The neural network may be trained as a conventional feed-forward back-propagation neural network. The value at the output is a score that signifies how likely it is that the particular linguistic feature is present. For example, acoustic model 316 may include a neural network trained to output a score indicating whether its input acoustic feature vector corresponds to a linguistic feature, such as a period. Additional neural networks may be trained for additional linguistic features (e.g., quotation marks, exclamation marks, commas, invisible phrasal boundaries). The scores from the neural networks are assumed to have a Gaussian distribution, which allows acoustic model 316 to convert the scores to a probability using the conventional Gaussian density function.

[0048] Fig. 6 is a flow chart illustrating methods for estimating a probability that one of the linguistic features is present at the location corresponding to the acoustic and lexical feature vector. Language model 315 estimates, based on the lexical vectors, the probability of a particular lexical vector corresponding to a boundary (act 601). Acoustic model 316 estimates the probability of the potential linguistic features occurring (act 602). Finally, statistical framework 303

combines the probabilities output from the language model 315 and the acoustic model 316 to generate the final probability that the linguistic feature is present (act 603). Statistical framework 303 may estimate this probability using maximum likelihood estimation (MLE) techniques. MLE techniques are well known in the art.

CONCLUSION

[0049] As described herein, a linguistic segmentation tool 115 generates linguistic features for a transcribed document. The linguistic features are associated with the original document as meta-information that enriches the content of the document.

[0050] The foregoing description of preferred embodiments of the invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. Moreover, while a series of acts have been presented with respect to Figs. 5 and 6, the order of the acts may be different in other implementations consistent with the present invention.

[0051] Certain portions of the invention have been described as software that performs one or more functions. The software may more generally be implemented as any type of logic. This logic may include hardware, such as application specific integrated circuit or a field programmable gate array, software, or a combination of hardware and software.

[0052] No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article "a" is intended to include one or more items. Where only one item is intended, the term "one" or similar language is used.

[0053] The scope of the invention is defined by the claims and their equivalents.